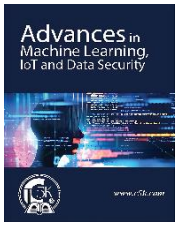
Advances in Machine Learning, IoT and Data Security
"Volume 1, Issue 2, Year 2025"
website: https://www.c5k.com



Research Article

Big Data Analytics and Its Usage on Financial Fraud Detection in the USA

Md Hossain Jamil[1], Arif Hossen[2], Shafiqul Islam Talukder[3], Yeasin Arafat[4], and Hasan Mahmud Sozib[5,*]

[1]*Department of Business Administration, Humphreys University, Stockton, CA 95207, USA; Hu0111561@student.humphreys.edu*

[2] *Department of Business Administration, International American University, Los Angeles, CA 90010, USA; arifhossen4295@gmail.com*

[3] *Department of Computer Science, Westcliff University, Irvine, CA 92614, USA; shafiqul.cse2017@gmail.com*

[4] *Department of Business Administration, Westcliff University, Irvine, CA 92614, USA; y.arafat.570@westcliff.edu*

[5] *Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Tejgaon, Dhaka-1208, Bangladesh; sozib2019@gmail.com*

**Corresponding Author: sozib2019@gmail.com*

ARTICLE INFO

Article history:
05 Jan 2025 (Received)
21 Feb 2025 (Accepted)
28 Feb 2025 (Published Online)

Keywords:
Big Data Analytics; Financial Fraud; Fraud Detection; Machine Learning; Risk Management; USA; Financial Services; Data Privacy.

ABSTRACT

Big data analytics has emerged as a transformative tool in the financial services industry, particularly in the United States, where institutions manage trillions of dollars in daily transactions. This study explores how financial institutions leverage big data analytics for risk management, with a specific focus on fraud detection and prevention. By integrating advanced technologies such as machine learning and artificial intelligence, big data analytics enables the real-time processing of vast datasets to uncover hidden patterns, identify anomalies, and predict potential threats. Traditional fraud detection methods often fail to address the growing complexity and sophistication of financial crimes. In contrast, machine learning models like Logistic Regression, Decision Trees, and Random Forests provide robust solutions by offering enhanced predictive accuracy and adaptability to evolving fraud tactics. This study examines a dataset comprising demographic, transactional, and geographical features, which are analyzed using machine learning algorithms. In order to guarantee fair and reliable fraud detection systems, the report emphasizes the need to strike a balance between regulatory compliance and technical improvements. The results highlight how crucial it is to include big data analytics into financial risk management plans in order to improve operational security and client confidence. To further increase the effectiveness of fraud detection, future research should concentrate on improving machine learning models, correcting biases, and investigating cutting-edge technologies like blockchain. This study confirms that big data analytics is an essential part of the continuous development of financial security and risk mitigation in the digital age, in addition to being a potent instrument for preventing fraud. Case studies from leading U.S. financial institutions, including JPMorgan Chase and PayPal, illustrate the real-world applications of big data in combating fraud. By integrating diverse data sources and leveraging advanced analytic techniques, these organizations have achieved notable reductions in fraudulent activities. The study concludes that big data analytics is not only a cornerstone of innovation and efficiency but also an essential component of modern risk management strategies. Future research should focus on addressing implementation challenges and exploring emerging technologies like blockchain to further enhance fraud detection capabilities.

DOI: https://doi.org/10.103/xxx @ 2025 Advances in Machine Learning, IoT and Data (AMLID), C5K Research Publication

1. Introduction

The term "big data" emerged in the early 2000s, but its conceptual roots trace back to the 1960s and 1970s when the first data storage and management systems were developed. The advent of relational databases in the 1980s marked a turning point, enabling organizations to store and retrieve large volumes of data efficiently (Codd, 1970). The real explosion of big data came with the rise of the internet, social media, and advanced computing technologies in the 21st century. Companies like Google and Amazon pioneered data-driven decision-making, leveraging massive datasets to predict user behavior and optimize services (Mayer-Schönberger & Cukier, 2013). Big data is categorized into three primary types: structured, unstructured, and semi-structured data. Structured data refers to highly organized information, such as transaction records and spreadsheets, stored in relational databases (George et al., 2014). Unstructured data includes text, images,

*Corresponding author: sozib2019@gmail.com (Hasan Mahmud Sozib)

Cite: Md Hossain Jamil, Arif Hossen, Shafiqul Islam Talukder, Yeasin Arafat, and Hasan Mahmud Sozib (2025). Big Data Analytics and Its Usage on Financial Fraud Detection in the USA. *Advances in Machine Learning, IoT and Data Security,* 1(2), pp. 1-12.

videos, and social media content, which lack a predefined format but offer valuable insights when analyzed using advanced tools (Gandomi & Haider, 2015). Semi-structured data, such as JSON and XML files, falls between these categories, combining elements of both structured and unstructured formats (Sagiroglu & Sinanc, 2013).

The versatility of big data analytics extends beyond finance, impacting various sectors such as healthcare, retail, manufacturing, and transportation. In healthcare, big data enables predictive analytics to improve patient outcomes and optimize resource allocation (Raghupathi & Raghupathi, 2014). Retailers use big data to personalize customer experiences, forecast demand, and streamline supply chains (Chaffey et al., 2019). Similarly, manufacturers leverage data analytics to enhance production efficiency and reduce downtime through predictive maintenance (Manyika et al., 2011). In transportation, big data facilitates route optimization, traffic management, and autonomous vehicle development (Zhang et al., 2011). These applications underscore the transformative potential of big data analytics across diverse domains, laying the foundation for its adoption in financial services.

Big data analytics offers numerous benefits to financial institutions, ranging from operational efficiency to enhanced customer experiences. By analyzing historical and real-time data, banks and financial firms can identify patterns and trends that inform strategic decisions (Roxburgh et al., 2011). For instance, personalized product recommendations and targeted marketing campaigns are powered by big data, fostering customer loyalty and satisfaction (Chambers & Dinsmore, 2015). Risk management is another critical area where big data analytics proves invaluable. Predictive modeling and machine learning algorithms enable institutions to assess creditworthiness, detect anomalies, and forecast potential risks with unprecedented accuracy (Ngai et al., 2011). Furthermore, big data enhances regulatory compliance by automating reporting processes and ensuring adherence to complex financial regulations (Olaiya et al., 2025).

Fraud detection and prevention represent one of the most significant applications of big data in the financial sector. Traditional fraud detection methods relied on rule-based systems that often failed to identify sophisticated schemes(Hancock & Khoshgoftaar, 2021). In contrast, big data analytics leverages machine learning and AI to analyze vast datasets, uncover hidden patterns, and detect fraudulent activities in real time (Dicuonzo et al., 2019). Big data analytics is indispensable in risk management and fraud detection, particularly for financial institutions (Politou et al., 2019). The financial services industry in the United States holds a crucial position in the global economy, managing trillions of dollars in daily transactions. In 2023, the average daily trading volume for equities reached 11.0 billion shares, equivalent to approximately $2.8 trillion(Weh et al., 2025). Similarly, the Federal Reserve's Fedwire Funds Service reported an average daily transfer value of $4.7 trillion in November 2025(Mohammadian Amiri & Esfahanipour, 2025). These figures highlight the significant scale of transactions handled by U.S. financial institutions (Aldasoro et al., 2020).

Case studies from U.S. financial institutions highlight the effectiveness of big data in combating fraud. For example, JPMorgan Chase employs advanced analytics to monitor transaction data and identify suspicious activities, resulting in significant reductions in financial fraud (Westerman et al., 2014). Similarly, PayPal uses big data and machine learning to enhance its fraud detection systems, ensuring secure transactions for millions of users worldwide (Kshetri, 2016). The implications of big data analytics in fraud and risk management are profound. By integrating diverse data sources, such as transaction records, social media activity, and geolocation data, financial institutions can create comprehensive risk profiles and respond to threats proactively (Rawat et al., 2019). This capability not only reduces financial losses but also strengthens trust and confidence among customers. Big data analytics has become a cornerstone of innovation and efficiency in the U.S. financial services sector. Its ability to process and analyze massive datasets enables institutions to enhance risk management, optimize operations, and deliver personalized services(Nassar & Kamal, 2021). The application of big data in fraud detection and prevention exemplifies its transformative potential, offering robust solutions to mitigate financial crimes. As technology continues to evolve, the strategic use of big data will remain pivotal in addressing emerging challenges and driving sustainable growth in the financial industry (Himeur et al., 2023; Kshetri, 2016).

This study aims to explore the use of big data in US financial institutions for risk management, specifically in fraud detection and prevention. The dataset, which includes customer demographics, transaction details, and fraud indicators, enables the application of big data techniques to identify and mitigate financial risks. Through data preprocessing, feature engineering, and exploratory analysis, patterns of fraudulent behavior are uncovered. Machine learning models trained on this data enable real-time fraud detection, assigning risk scores and flagging suspicious activities. This process empowers financial institutions to proactively address fraud, safeguard against financial losses, and enhance operational security. By continuously monitoring and updating models with new data, institutions ensure their fraud detection systems remain adaptive to emerging threats.

2. Literature Review

Big Data refers to the massive volume of structured and unstructured data generated from various sources, including transactions, social media, and customer interactions. Analytics involves applying statistical and computational techniques to extract meaningful insights from the data. Big data became a mainstream concept around 2010, driven by Gartner's "Three Vs" framework: volume, velocity, and variety (Laney, 2001). This framework underscores the massive scale of data, its rapid generation, and the diverse formats it encompasses. Since then, continuous advancements in cloud computing, artificial intelligence (AI), and machine learning have propelled big data analytics into becoming a cornerstone of modern business strategies. These advancements enable the processing of immense datasets with enhanced accuracy and efficiency, facilitating transformative applications across industries, including finance (Ravi & Kamaruddin, 2017).

Financial fraud presents a serious threat to the financial sector with its growing complexity and scope endangering the stability of the world economy. As financial systems become more complex and transactions increasingly move to digital platforms, the sector faces heightened risks, including fraud, cybercrime, and operational inefficiencies (Aldasoro et al., 2020). Traditional fraud detection methods, such as manual audits and rule-based systems, have proven insufficient in dealing with the volume and sophistication of modern financial crimes. In response, financial institutions have turned to big data analytics, which enables them to process vast amounts of structured and unstructured data in real-time, uncover hidden patterns, and identify potential threats more accurately and quickly than ever before (Nobanee et al., 2021). Advanced machine learning (ML) and artificial intelligence (AI) algorithms facilitate predictive analytics, enhancing the ability to detect and prevent fraud proactively (Shoetan et al., 2025). Moreover, big data analytics enables institutions to incorporate diverse data sources, including transactional data, behavioral patterns, and unstructured data from social media, providing a more holistic approach to risk assessment (Mhlanga, 2025). By leveraging advanced analytic techniques, big data analytics reduces false positives for legitimate transactions and false negatives for fraudulent transactions, ensuring that genuine transactions are not mistakenly flagged as fraud and that actual fraudulent activities are accurately detected. This proactive approach allows for timely interventions, reducing the impact of fraud on both the institution and its customers (Gambacorta et al., 2025).

The application of big data analytics in fraud detection is multi-faceted, involving techniques such as Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR), and Probabilistic Neural Network (PNN) (Ravisankar et al., 2011). For instance, machine learning models trained on large datasets can identify subtle fraudulent patterns that would be undetectable through traditional methods. These models continuously improve by learning from new data, ensuring adaptability to evolving fraud tactics (Shoetan et al., 2025).

The effectiveness of big data analytics in fraud detection is further improved by the development of machine learning and artificial intelligence (AI). AI-powered big data analytics can handle enormous datasets at previously unheard-of speeds, allowing financial institutions to constantly adjust to emerging fraud tactics (Ali et al., 2022). Because financial fraudsters regularly alter their strategies to evade conventional detection methods, this flexibility is essential. Consequently, big data analytics lowers the time and expense involved in looking into fraudulent transactions while simultaneously increasing the accuracy of fraud detection (Martins & Fonkem, 2025). Additionally, big data analytics makes it easier to integrate many data sources, which improves the identification of intricate fraud schemes involving coordinated actions across various platforms. Financial fraud is increasingly occurring through a variety of channels, including e-commerce platforms, mobile payments, and online banking(VenkateswaraRao et al., 2023). Big data analytics makes it possible for these data streams to be seamlessly integrated, giving organizations a thorough understanding of consumer behavior and enabling them to identify fraudulent activity that may occur across several platforms and geographical areas (Shalhoob et al., 2025).

Despite the significant advantages that big data analytics offers, its implementation is not without challenges. One of the primary concerns is data privacy and security, as financial institutions must manage sensitive customer information while complying with stringent regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) (Sharma et al., 2025). These regulations aim to protect customers' data and ensure transparency in how it is used. Moreover, there is growing concern over the potential for algorithmic bias in machine learning models used for fraud detection. While AI and machine learning can uncover patterns in vast datasets, these models are only as good as the data they are trained on (Mittal, 2013). If biased or incomplete data is used, the algorithms may produce inaccurate results, leading to unfair outcomes, such as disproportionate targeting of certain demographic groups (Verma, 2019). Financial institutions must therefore be vigilant in ensuring that their big data models are both accurate and ethical, balancing the need for security with respect for individual privacy and fairness(Al-Hashedi & Magalingam, 2021).

The widespread use of big data analytics in the financial sector is revolutionizing the way institutions approach risk management. By improving fraud detection and

providing more accurate risk assessments, big data helps reduce financial losses and builds trust with customers (Thennakoon et al., 2019). With the ability to detect fraud in real time, institutions can offer a safer and more seamless experience for their clients, enhancing customer satisfaction and loyalty (Fernando et al., 2018). Moreover, big data allows financial institutions to stay ahead of emerging risks by continuously analyzing new data, ensuring that their fraud detection models remain adaptive and responsive to evolving threats (Bello et al., 2025).

The literature review emphasizes the definition and importance of big data. The section also discovers the role of big data analysis in different fields. Fraud detection and its revolution through the years have also been explained. Financial fraud has been a problem for businesses for a long time. Big data analysis is becoming popular for fraud detection in the USA as well as other countries. As more case studies demonstrate the success of big data in fraud detection and risk management, it becomes clear that this technology is critical to the future of the financial services industry. This study explores how USA financial institutions leverage big data analytics for fraud detection.

3. Data Collection and Preprocessing

To achieve the objectives of this study, a fraud detection dataset provided by Git Hub was utilized(*Github*). The dataset included key metrics such as age, gender, ZIP code, merchant ZIP code (zipMerchant), transaction category, transaction amount, transaction count, and fraud labels. By analyzing these variables, patterns, and anomalies associated with fraudulent activities were identified. Advanced analytics techniques, including machine learning algorithms and statistical modeling, were employed to explore correlations and trends. For instance, age and gender distributions were assessed to determine demographic factors linked to fraud, while spatial analysis of ZIP codes and merchant locations helped uncover geographic hotspots for suspicious activities. The dataset's comprehensive nature enabled a detailed examination of transaction categories and amounts, revealing insights into high-risk behaviors and providing actionable intelligence for fraud prevention strategies.

3.1. Data Preprocessing and Cleaning

The Table I dataset underwent data preprocessing, including handling missing values, converting categorical variables, and transforming features for compatibility with machine learning models. All categorical variables were assigned unique integers for numerical compatibility. The 'Amount' feature was converted to float, ensuring numerical precision.

Table 1. Features of the Dataset

Feature	Feature Description
Age	Age of the transaction initiator or cardholder
Gender	Gender of the transaction initiator or cardholder
Transaction Amount	Amount of money involved in the transaction
Transaction Count	Number of transactions associated with the event
Transaction Category	Type of transaction (e.g., health, transportation)
ZIP Code	Postal code of the transaction location or cardholder
Merchant ZIP Code	Postal code of the merchant involved in the transaction
Fraud	Indicating whether the transaction was fraudulent or not

Table 1 illustrates some key features of the dataset. The features include demographic information like age and gender. It also includes transaction behavior-related information like transaction amount, transaction count, and transaction category. Furthermore, some geographical information of cardholders or transaction initiators is also captured by the dataset. It includes the Zip code, and merchant zip code. Overall, table 1 gives an overview of the information captured within the dataset, which is used for the analysis and modeling of fraudulent transactions.

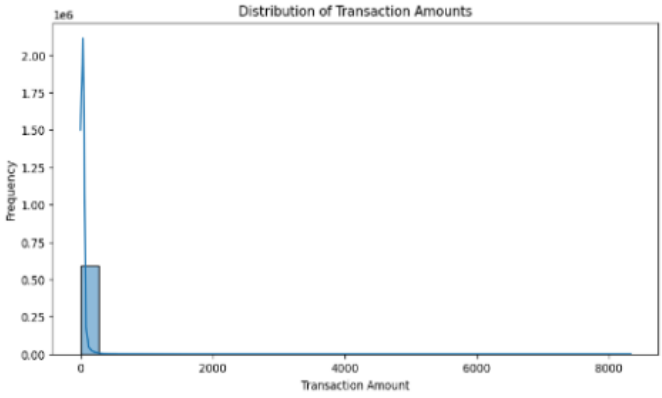


Fig. 1(A)

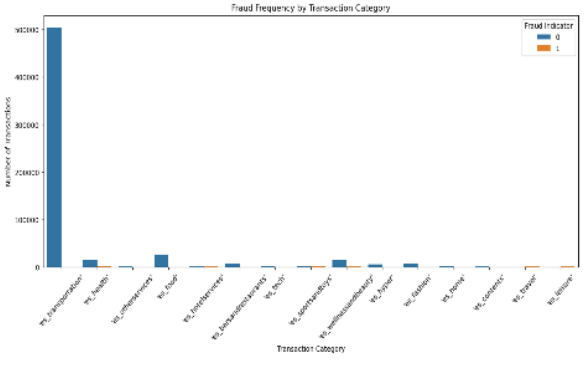


Fig. 1(B)

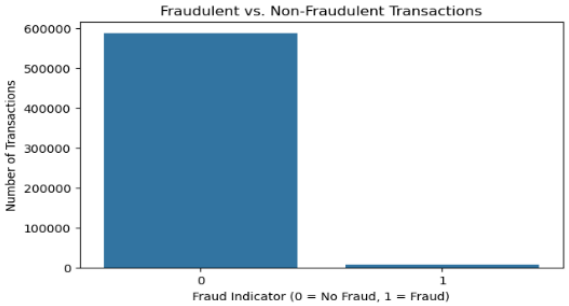


Fig. 1(C)

Fig. 1(D)

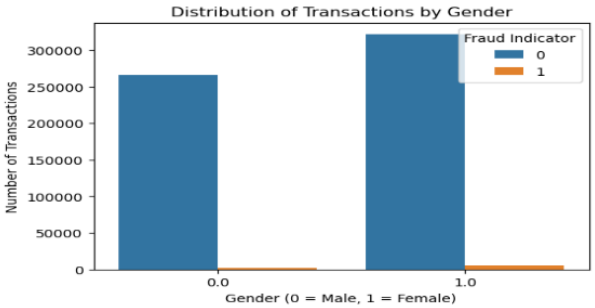
Fig.1. Distribution analysis of key metrics.

Figure 1 exhibits several exploratory data analysis (EDA) plots related to fraud detection in financial transactions. They visualize the distribution and relationships of key variables with the target variable (fraudulent vs. non-fraudulent transactions).

The first histogram in Figure 1(A), shows the distribution of transaction amounts. The majority of transactions are concentrated in the lower range, with the frequency gradually decreasing as the transaction amount increases. The next graph in Figure 1(B) displays the "Fraud Frequency by Transaction Category". The horizontal axis represents the different transaction categories. The vertical axis represents the "Number of Transactions". The scale is marked in increments of 100,000, ranging from 0 to 500,000. The bars in the chart represent the number of transactions in each category. The graph shows that the majority of transactions in each category are non-fraudulent, with the blue bars being significantly taller than the orange bars for each category. The next bar chart, Figure 1(C), shows the distribution of fraudulent and non-fraudulent transactions across different age groups. The number of fraudulent transactions is generally lower than non-fraudulent transactions in all age groups and also shows a decreasing trend with increasing age groups. It also indicates that fraudulent activity is less common overall compared to non-fraudulent transactions, and the frequency of both fraudulent and non-fraudulent transactions tends to decrease with increasing age. The following bar chart in Figure 1(D) shows the distribution of fraudulent and non-fraudulent transactions by gender. There are more transactions (both fraudulent and non-fraudulent) for females (gender = 1) compared to males (gender = 0). This suggests that fraudulent activity is significantly higher for the female gender in this dataset. Finally, the last bar chart illustrates the number of fraudulent and non-fraudulent transactions. The vast majority of transactions are non-fraudulent (indicated by a value of 0), while fraudulent transactions (indicated by a value of 1) are significantly less frequent.

4. Models for predictions

Machine learning employs various predictive models to analyze data patterns and forecast future events. This study u
Logistic
Each m
and app



4.1. Log

Logistic
predicti
particul
where the outcome can be one of two possible values (e.g., yes/no, fraud/not fraud). The model estimates the probability of the event based on a set of input features using a sigmoid function, which maps the input values to a probability between 0 and 1 (Itoo et al., 2021; Song et al., 2021).

The formula for logistic regression is given by:

$$P(Y = 1/X) = \frac{1}{(1+e-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)} \quad \ldots\ldots\ldots (01)$$

Where,

$P(Y = 1/X)$ is the probability of the outcome, β represents the coefficients, and X represents the input features (Das, 2025).

4.2. Decision Trees

Decision Trees are a type of supervised learning algorithm that creates a tree-like model of decisions and their possible consequences. They work by recursively partitioning the data based on the values of the input features, creating a series of if-then-else rules. The resulting tree can then be used to predict the outcome for new data points by following the path down the tree

based on their feature values (Costa & Pedreira, 2023; Xu et al., 2023).

4.3. Random Forest

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees. It works by creating a large number of decision trees, each trained on a different subset of the data and using a random subset of features. The final prediction is made by aggregating the predictions of all the individual trees, often by taking a majority vote or averaging their predictions. Random Forest is known for its high accuracy and robustness to overfitting (Lin & Jiang, 2021; Wu & Chang, 2025).

So, Logistic Regression provides a probabilistic prediction, Decision Trees offer a clear and interpretable set of rules, and Random Forest leverages the power of multiple trees to achieve high accuracy(Liu, 2021; Mehbodniya et al., 2021).

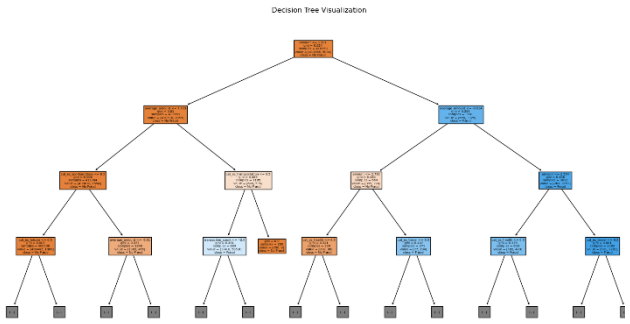


Fig. 2. Decision Tree Visualization.

A decision tree is a flowchart-like structure where each internal node represents a feature (or attribute), each branch represents a decision rule, and each leaf node represents an outcome (in this case, whether a transaction is classified as fraudulent or not) (Yin et al., 2021). The tree is built by splitting the dataset based on feature values to create branches that lead to different outcomes.

Before constructing the decision tree, the dataset undergoes preprocessing, which includes:

4.3.1. Data Processing

- **Handling Missing Values**: Any missing data points in the dataset are addressed to ensure that the analysis is based on complete information.
- **Converting Categorical Variables**: Categorical features (like gender or transaction category) are transformed into numerical formats by assigning unique integers. This step is essential for compatibility with machine learning algorithms.
- **Transforming Features**: Features such as transaction amounts are converted to appropriate numerical types (e.g., float) to maintain precision during calculations.

4.3.2. Feature Selection

The decision tree algorithm evaluates all features in the dataset to determine which ones best separate the classes (fraudulent vs. non-fraudulent transactions). This involves:

- **Calculating Impurity Measures**: The algorithm uses metrics like Gini impurity or entropy to assess how well a feature distinguishes between classes.
    - **Gini Impurity** is calculated using the formula:

$$Gini = 1 - \sum(P_i^2) \quad \ldots\ldots\ldots (02)$$

Where P_i is the proportion of each class in the node. A lower $Gini$ impurity value indicates better separation between classes. Moreover, it goes through other processes like splitting criteria, recursive splitting, and leaf node outcomes (Charbuty & Abdulazeez, 2021).

Figure 2 shows that the root node consists of the feature: amount $<= 1.99$. The initial decision point categorizes transactions based on whether their amount is less than or equal to 1.99. This suggests that lower transaction amounts are treated differently in the classification process.

In the subsequent branches, the tree evaluates the average_amount at thresholds of 7.73 and -0.01, indicating that this feature is significant in distinguishing between classes. For average_amount $<= 7.73$, further splits based on categorical variables like cat_es_sportsandtoys and cat_es_leisure are made, suggesting a detailed analysis of transaction characteristics. For average_amount > 7.73, the decision tree considers other categorical variables, such as cat_es_transportation, indicating a shift in focus for higher transaction amounts. The leaf nodes indicate classifications, such as class: 0, which typically denotes legitimate transactions. This highlights how various paths through the tree lead to different classifications based on prior decisions. The presence of truncated branches (e.g., "truncated branch of depth 34") indicates that there are additional levels of complexity and decision-making not fully represented in this overview. This suggests a comprehensive model capable of capturing intricate patterns in the data.

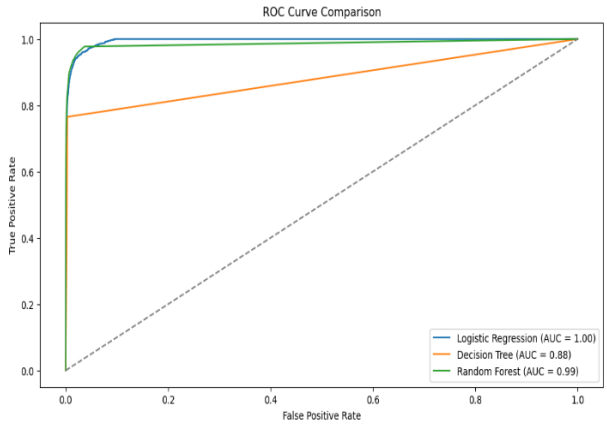


Fig. 3(A)

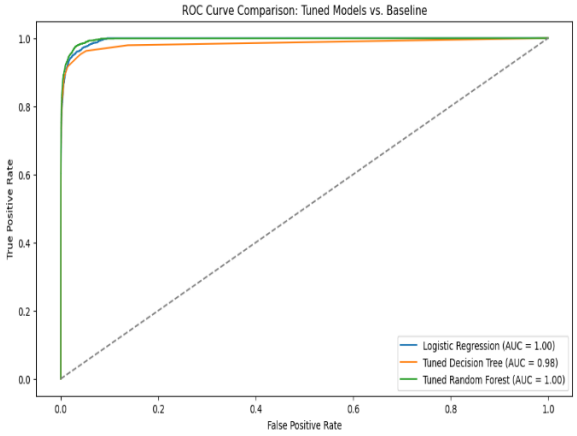


Fig. 3(B)

Fig. 3. ROC curve comparison

Figure 3 presents two ROC (Receiver Operating Characteristic) curves. ROC curves are a common tool used in machine learning to visualize and compare the performance of binary classification models. They plot the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The area under each curve (AUC) represents the overall performance of the models. A higher AUC generally indicates better performance.

ROC Curve in Figure 3(A) compares the performance of three models: Logistic Regression, Decision Tree, and Random Forest. Logistic regression appears to have perfect performance, with an AUC of 1.00. Decision Tree (AUC = 0.88) also performs reasonably well, but not as perfectly as Logistic Regression. Random Forest (AUC = 0.99) shows excellent performance, just slightly below the perfect performance of Logistic Regression.

The ROC Curve in Figure 3(B) compares the performance of the same three models (Logistic Regression, Decision Tree, and Random Forest) after they have been tuned or optimized. It also includes a baseline model. The performance of the tuned Logistic Regression model (AUC = 1.00) remains perfect. The tuned Decision Tree model (AUC = 0.98) shows a slight performance improvement compared to the untuned version. The tuned Random Forest model (AUC = 1.00) also achieved perfect performance. The baseline model represents a simple classifier that always predicts the same class. It serves as a reference point for evaluating the performance of the tuned models.

Hence, tuning led to improved performance for the Decision Tree model and perfect performance for the Random Forest model.

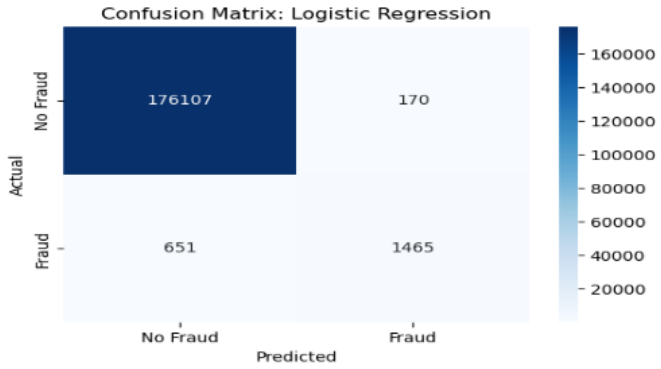


Fig. 4(A). Confusion matrices of Logistic Regression.

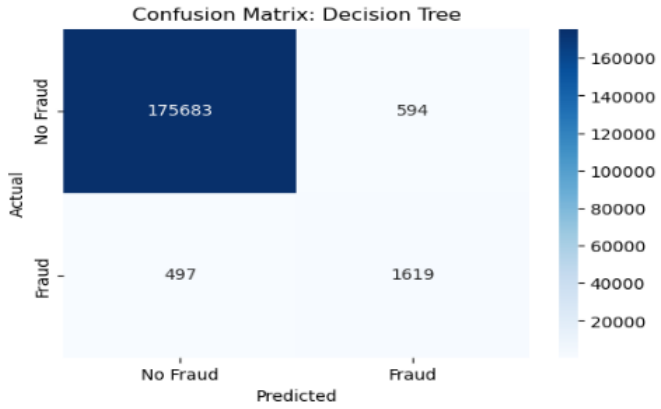


Fig. 4(B). Confusion matrices of Decision Tree.

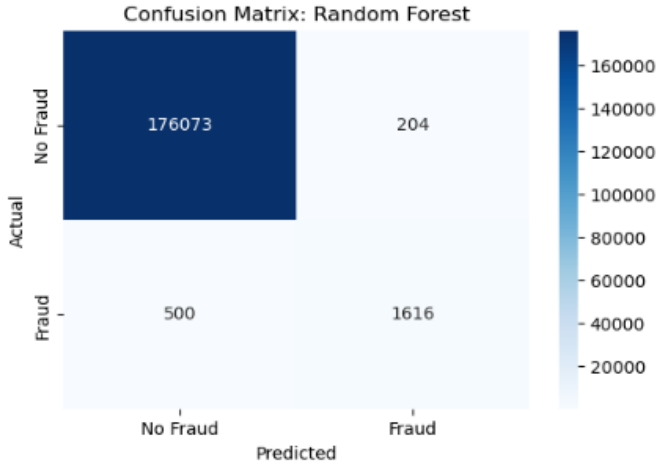


Fig. 4(C). Confusion matrices of Random Forest.

Figure 4 presents three confusion matrices. A confusion matrix in Figures is a visualization tool used in machine learning to evaluate the performance of classification models. It helps understand how well a model can correctly classify instances into different categories, especially in cases where the classes might be imbalanced. Each matrix has two rows and two columns:

1. Actual: Represents the true class labels (whether a transaction is Fraud or No Fraud).
2. Predicted: Represents the class labels predicted by the model.

The confusion matrices for the Logistic Regression, Decision Tree, and Random Forest classifiers, respectively, in a binary classification task for identifying "Fraud" and "No Fraud" instances are shown in Figures 4(A), 4(B), and 4(C). The performance of the models is shown in each matrix in terms of false positives, false negatives, true positives, and true negatives. 176,107 "No Fraud" and 1,465 "Fraud" cases were accurately identified using Logistic Regression in Figure 4(A), however, 651 fraudulent cases were incorrectly classified as "No Fraud" and 170 non-fraudulent cases as "Fraud." This implies that Logistic Regression favors high precision for "No Fraud" instances and has trouble detecting fraud. The Decision Tree model, shown in Figure 4(B), successfully detected 175,683 "No Fraud" cases and increased the number of "Fraud" occurrences detected to 1,619 cases. Comparing this improvement to Logistic Regression, however, resulted in more false positives (594 "No Fraud" instances incorrectly categorized as "Fraud"). The trade-off suggests that the Decision Tree detects fraud a little more aggressively. The Random Forest model, which accurately predicted 176,073 "No Fraud" instances and 1,616 "Fraud" cases, obtained a nearly ideal equilibrium, as shown in Figure 4(C). While retaining good fraud detection performance, it drastically decreased false positives (204), incorrectly identifying just 500 fraudulent instances as "No Fraud." The best model for the specified classification issue is Random Forest, which performs better overall and strikes a better balance between sensitivity and specificity. The trade-offs among the models' accuracy, recall, and misclassification rates are highlighted by these matrices.

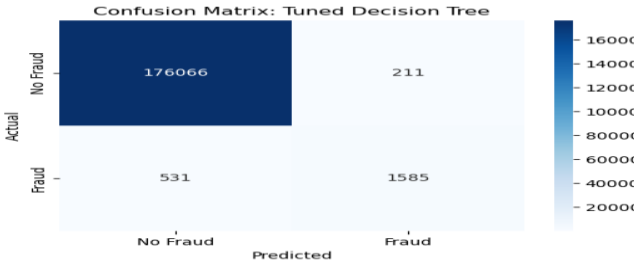


Fig. 5(A). Confusion matrix for Tuned Decision Tree.

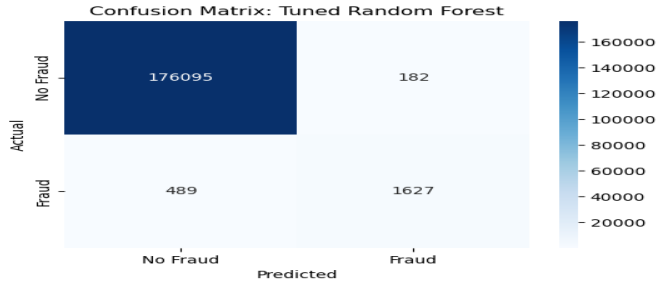


Fig. 5(B). Confusion matrices for Tuned Random Forest.

Figure 5 presents two confusion matrices Figure 5(A) and Figure 5(B). In the figures, the models are trained to distinguish between legitimate transactions ("No Fraud") and fraudulent ones ("Fraud"). Both models appear to have high accuracy, with the Tuned Decision Tree likely having a slight edge due to fewer false positives. The Tuned Decision Tree seems to have higher precision, meaning it has a lower proportion of false positives among its "Fraud" predictions. It also seems to have slightly better recall, meaning it correctly identifies a higher percentage of actual fraudulent transactions. Both models have very high specificity, indicating they are excellent at correctly identifying legitimate transactions. In fraud detection, false positives can lead to inconvenience for customers and potential damage to the bank's reputation. Therefore, high precision is crucial.

5. Result and Analysis

Following the data processing and prediction model selection via the ROC curve and the confusion matrix, the random forest and Decision tree out-standard other prediction models for studying the role of big data for financial fraud detection in the USA. After running several tests with different models a perfect heatmap of the dataset was analyzed. To shed light on the numerical features of the fraud detection analysis, Figure 6 pictures a heatmap for the dataset.

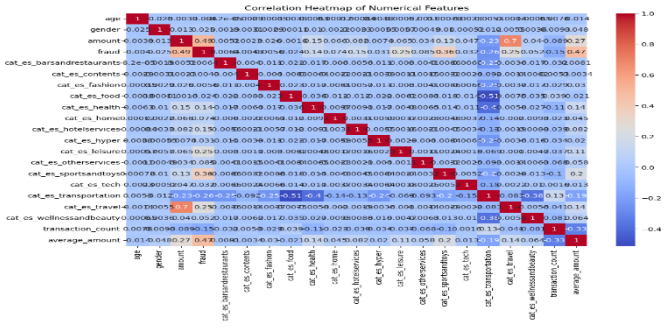


Fig. 6. Correlation Heatmap for numerical features.

With values ranging from -1 (strong negative correlation) to 1 (strong positive correlation), the correlation heatmap illustrates the connections between numerical characteristics. For instance, there is a moderately positive association (0.49) between "fraud" and "amount," suggesting that fraud may be associated

with larger transaction amounts. "cat_es_travel" and "cat_es_hotelservices" have the largest positive correlation (0.70), indicating a relationship between these expenditure categories. On the other hand, the correlation between "cat_es_transportation" and "cat_es_travel" is negative (-0.40), suggesting less overlap. Interestingly, "average_amount" has a negative correlation with "transaction_count" (-0.33) and a correlation with "amount" (0.47). These revelations aid in determining feature dependencies for more research.

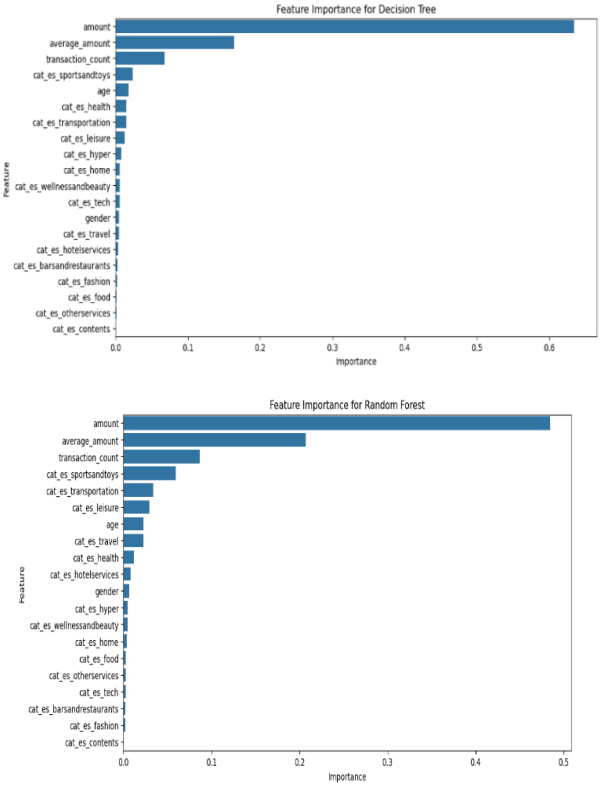


Fig. 7. Feature importance for Decision Tree and Random Forest

Figure 7 visualizes the relative importance of different features (variables) in predicting fraud in financial transactions. The importance is determined by how much each feature contributes to the decision-making process of the machine learning models used: A Decision Tree and a Random Forest. In the plot, the horizontal axis represents the importance of each feature. A higher value indicates greater importance while the vertical axis lists the different features considered by the models.

In case of the Decision Tree, transaction amount (amount) is the most important feature. This makes intuitive sense as large or unusual transactions are often associated with fraud. Average transaction amount (average_amount) is the second most important feature. This suggests that patterns in spending behavior are significant indicators of fraud. Transaction count (transaction_count) is the third most important feature. Frequent transactions within a short period might be a red flag. Various categories like transactions of sports and toys, health, transportation, etc., also play a role. These likely capture spending patterns in different domains that can be indicative of fraudulent activity.

For Random Forest, Transaction Amount (amount) remains the most important feature, reinforcing its significance. Average transaction amount (average_amount) is still highly important, indicating its consistent relevance across models. Transaction count (transaction_count) maintains its importance, suggesting its consistent predictive power. The order of importance for categories differs slightly between the two models, but many categories remain influential in both cases.

So, focusing on transaction amounts and spending patterns is crucial for fraud detection. Analyzing spending behavior across different categories can reveal suspicious activities. The similarity between the two models' feature importance rankings suggests that these features are robust indicators of fraud.

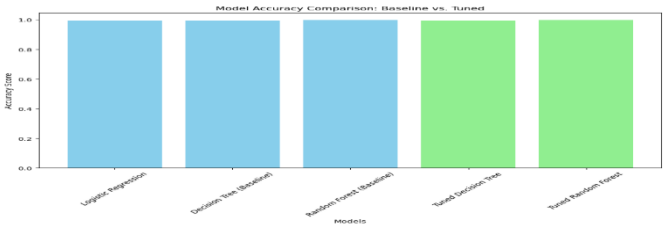


Fig. 8. Model Accuracy Comparison: Baseline vs. Tuned.

In figure 8, the X-axis represents the different models used for classification, viz., Logistic Regression, Decision Tree (Baseline), Random Forest (Baseline), Tuned Decision Tree, Tuned Random Forest while the Y-axis represents the accuracy score of each model, ranging from 0 to 1.

All three baseline models (Logistic Regression, Decision Tree, and Random Forest) show a similar level of accuracy, all around 1.0. This suggests that these models, in their initial state, are already performing quite well. Both the tuned Decision Tree and the tuned Random Forest models also show an accuracy of 1.0. This indicates that the tuning process did not result in a significant improvement in accuracy for these models. While accuracy is important, other metrics like precision, recall, and F1-score are crucial in fraud detection. Precision is important to minimize false positives (flagging legitimate transactions as fraudulent), and recall is crucial to minimize false negatives (missing actual fraudulent transactions).

6. Conclusion and Future Work

With its unmatched ability to analyze enormous volumes of data in real time, big data analytics has become a game-changer in the identification of financial crime. This study demonstrates how well machine learning models like Random Forest, Decision Trees, and Logistic Regression can detect fraudulent

transactions. Random Forest outperformed the others in terms of precision and recall, reducing false positives and negatives. These findings suggest that fraud detection systems may be greatly improved by sophisticated analytics, strengthening financial institutions' defenses against changing threats. Fraud prevention has significantly improved as a result of financial institutions' implementation of big data analytics. Financial fraud has been effectively decreased by the use of real-time transaction monitoring, anomaly detection, and predictive modeling, as demonstrated by case studies from top U.S. banks and digital payment platforms such as JPMorgan Chase and PayPal. Financial institutions are able to create a thorough risk profile for every client by utilizing information from several sources, including social media interactions, transaction records, and geolocation data. Proactive mitigation techniques and quicker fraud detection are made possible by this all-encompassing strategy.The capacity of big data analytics to develop in tandem with new fraud strategies is one of its main benefits. Conventional rule-based systems frequently fall short in identifying complex fraudulent schemes that take advantage of weaknesses in financial procedures. On the other hand, machine learning algorithms improve their fraud detection skills over time by continually learning from fresh data. This flexibility is essential as thieves create ever-more intricate plans to get around traditional security systems. Fraud detection frameworks are further strengthened by the application of deep learning, natural language processing, and real-time behavioral analysis, which guarantees that fraudulent acts are discovered before they result in substantial financial harm.

The application of big data analytics in financial fraud detection is not without difficulties, nevertheless, despite its revolutionary promise. Data security and privacy are two key issues. Strict laws like the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR), which require openness in data usage, must be followed by financial institutions. Maintaining high accuracy while protecting client privacy in fraud detection algorithms is a difficult balance that calls for constant monitoring. Moreover, the moral use of machine learning models is still a major concern. Unfair treatment of particular demographic groups due to algorithmic bias might result in regulatory attention and reputational issues. Building fair fraud detection systems requires addressing these biases using representative and varied training datasets. The computational expense of big data analytics is another major obstacle. Large dataset processing necessitates a significant amount of processing power and storage capacity, which raises operating costs. To effectively handle these issues, financial institutions need to make investments in scalable and reasonably priced technology, such as distributed computing frameworks and cloud-based solutions. Future studies should look for methods to improve fraud detection systems even further. Blockchain integration has the potential to increase transaction security and transparency. Because blockchain technology is decentralized, it may produce an unchangeable record of transactions, making fraudulent changes all but impossible. Furthermore, federated learning in which models are trained on decentralized data sources without disclosing private information could improve privacy while preserving the accuracy of fraud detection. To sum up, big data analytics is transforming the detection of financial fraud by providing cutting-edge instruments for instantly detecting and stopping fraudulent activity. Even while problems like algorithmic bias, data privacy, and computing costs still exist, they can be lessened with ongoing technical developments and legislative frameworks. Financial institutions may improve their fraud detection skills and create a more secure financial environment by adopting new technology and improving machine learning models. The smooth integration of big data analytics, ethical AI, blockchain, and cooperative industry efforts to successfully combat financial crimes is the key to the future of financial fraud prevention.

References

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402. https://doi.org/https://doi.org/10.1016/j.cosrev.2021.100402

Aldasoro, I., Gambacorta, L., Giudici, P., & Leach, T. (2020). Operational and cyber risks in the financial sector.

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences*, *12*(19), 9637.

Bello, H. O., Ige, A. B., & Ameyaw, M. N. (2025). Adaptive machine learning models: concepts for real-time financial fraud prevention in dynamic environments. *World Journal of Advanced Engineering Technology and Sciences*, *12*(02), 021-034.

Chaffey, D., Edmundson-Bird, D., & Hemphill, T. (2019). *Digital business and e-commerce management*. Pearson Uk.

Chambers, M., & Dinsmore, T. W. (2015). *Advanced analytics methodologies: Driving business value with analytics*. Pearson Education.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377-387.

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, *56*(5), 4765-4800.

Das, A. (2025). Logistic regression. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 3985-3986). Springer.

Dicuonzo, G., Galeone, G., Zappimbulso, E., & Dell'Atti, V. (2019). Risk management 4.0: The role of big data analytics in the bank sector. *International Journal of Economics and Financial Issues*, *9*(6), 40-47.

Fernando, Y., Chidambaram, R. R., & Wahyuni-TD, I. S. (2018). The impact of Big Data analytics and data security practices on service supply chain performance. *Benchmarking: An International Journal*, *25*(9), 4009-4034.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2025). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *Journal of Financial Stability*, 101284.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. In (Vol. 57, pp. 321-326): Academy of Management Briarcliff Manor, NY.

Github. https://raw.githubusercontent.com/atavci/fraud-detection-on-banksim-data/master/Data/synthetic-data-from-a-financial-payment-system/bs140513_032310.csv

Hancock, J. T., & Khoshgoftaar, T. M. (2021). Gradient Boosted Decision Tree Algorithms for Medicare Fraud Detection. *SN Computer Science*, *2*(4), 268. https://doi.org/10.1007/s42979-021-00655-z

Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., & Amira, A. (2023). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. *Artificial Intelligence Review*, *56*(6), 4929-5021. https://doi.org/10.1007/s10462-022-10286-2

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503-1511. https://doi.org/10.1007/s41870-020-00430-y

Kshetri, N. (2016). Big data's role in expanding access to financial services in China. *International journal of information management*, *36*(3), 297-308.

Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety, Application delivery strategies. *Stamford: META Group Inc*, *1*.

Lin, T.-H., & Jiang, J.-R. (2021). Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest. *Mathematics*, *9*(21).

Liu, X. (2021). Empirical Analysis of Financial Statement Fraud of Listed Companies Based on Logistic Regression and Random Forest Algorithm. *Journal of Mathematics*, *2021*(1), 9241338. https://doi.org/https://doi.org/10.1155/2021/9241338

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition*.

Martins, O., & Fonkem, B. (2025). Leveraging big data analytics to combat emerging financial fraud schemes in the USA: a literature review and practical implications. *World J Adv Res Reviews*, *24*, 17-43.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mehbodniya, A., Alam, I., Pande, S., Neware, R., Rane, K. P., Shabaz, M., & Madhavan, M. V. (2021). [Retracted] Financial Fraud Detection in Healthcare Using Machine Learning and Deep Learning Techniques. *Security and Communication Networks*, *2021*(1), 9293877. https://doi.org/https://doi.org/10.1155/2021/9293877

Mhlanga, D. (2025). The role of big data in financial technology toward financial inclusion. *Frontiers in big Data*, *7*, 1184444.

Mittal, A. (2013). Trustworthiness of big data. *International Journal of Computer Applications*, *80*(9).

Mohammadian Amiri, E., & Esfahanipour, A. (2025). A hybrid approach for a novel dynamic trading system to produce robust cryptocurrency portfolios. *Scientia Iranica*.

Nassar, A., & Kamal, M. (2021). Machine Learning and Big Data Analytics for Cybersecurity Threat Detection: A Holistic Review of Techniques and Case Studies. *Journal of Artificial Intelligence and Machine Learning in Management*, *5*(1), 51-63. https://journals.sagescience.org/index.php/jamm/article/view/97

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559-569.

Nobanee, H., Dilshad, M. N., Al Dhanhani, M., Al Neyadi, M., Al Qubaisi, S., & Al Shamsi, S. (2021). Big Data applications the banking sector: A bibliometric analysis approach. *Sage Open*, *11*(4), 21582440211067234.

Olaiya, O. P., Cynthia, A. C., Usoro, S. O., Obani, O. Q., Nwafor, K. C., & Ajayi, O. O. (2025). The impact of big data analytics on financial risk management. *Int. J. Sci. Res. Arch*, *12*(2), 821-827.

Politou, E., Alepis, E., & Patsakis, C. (2019). Profiling tax and financial behaviour with Big Data under the GDPR. *Computer law & security review*, *35*(3), 306-329.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*, 1-10.

Ravi, V., & Kamaruddin, S. (2017). Big data analytics enabled smart financial services: opportunities and challenges. Big Data Analytics: 5th International Conference, BDA 2017, Hyderabad, India, December 12-15, 2017, Proceedings 5,

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision support systems*, *50*(2), 491-500.

Rawat, D. B., Doku, R., & Garuba, M. (2019). Cybersecurity in big data era: From securing big data to data-driven security. *IEEE Transactions on Services Computing*, *14*(6), 2055-2072.

Roxburgh, C., Lund, S., & Piotrowski, J. (2011). Mapping global capital markets 2011. *McKinsey Global Institute*, *201*(1), 1-38.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. 2013 international conference on collaboration technologies and systems (CTS),

Shalhoob, H., Halawani, B., Alharbi, M., & Babiker, I. (2025). The impact of big data analytics on the detection of errors and fraud in accounting processes. *RGSA: revista de gestão social e ambiental*, *18*(1).

Sharma, K., Kumar, P., & Özen, E. (2025). Ethical Considerations in Data Analytics: Challenges, Principles, and Best Practices. In *Data Alchemy in the Insurance Industry: The Transformative Power of Big Data Analytics* (pp. 41-48). Emerald Publishing Limited.

Shoetan, P. O., Oyewole, A. T., Okoye, C. C., & Ofodile, O. C. (2025). Reviewing the role of big data analytics in

financial fraud detection. *Finance & Accounting Research Journal*, *6*(3), 384-394.

Song, X., Liu, X., Liu, F., & Wang, C. (2021). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International journal of medical informatics*, *151*, 104484.

Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019). Real-time credit card fraud detection using machine learning. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence),

VenkateswaraRao, M., Vellela, S., V. R, B., Vullam, N., Sk, K. B., & R, D. (2023, 17-18 March 2023). Credit Investigation and Comprehensive Risk Management System based Big Data Analytics in Commercial Banking. 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS),

Verma, S. (2019). Weapons of math destruction: how big data increases inequality and threatens democracy. *Vikalpa*, *44*(2), 97-98.

Weh, R., Westerholm, P. J., Wilkens, M., & Yao, J. (2025). Liquidity provision and trading skill: Evidence from mutual funds' daily transactions. *Review of Financial Economics*, *42*(2), 206-238. https://doi.org/https://doi.org/10.1002/rfe.1196

Westerman, G., Bonnet, D., & McAfee, A. (2014). *Leading digital: Turning technology into business transformation*. Harvard Business Press.

Wu, Y.-c., & Chang, Y.-l. (2025). Ransomware detection on linux using machine learning with random forest algorithm. *Authorea Preprints*.

Xu, B., Wang, Y., Liao, X., & Wang, K. (2023). Efficient fraud detection using deep boosting decision trees. *Decision Support Systems*, *175*, 114037. https://doi.org/https://doi.org/10.1016/j.dss.2023.114037

Yin, L., Lin, X., Liu, J., Li, N., He, X., Zhang, M., Guo, J., Yang, J., Deng, L., Wang, Y., Liang, T., Wang, C., Jiang, H., Fu, Z., Li, S., Wang, K., Guo, Z., Ba, Y., Li, W., . . . Clinical Outcome of Common Cancers, G. (2021). Classification Tree–Based Machine Learning to Visualize and Validate a Decision Tool for Identifying Malnutrition in Cancer Patients. *Journal of Parenteral and Enteral Nutrition*, *45*(8), 1736-1748. https://doi.org/https://doi.org/10.1002/jpen.2070

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1624-1639.